In vitro veritas 2013; 14:137-146 ISSN: 1697-5421

Document docent

Associació Catalana de Ciències de Laboratori Clínic
Secció d'Estadística i Metrologia[1]

Estadística descriptiva aplicada a variables discretes

Preparat per:
Henar Valbuena Parralejo, Marta Mosquera Parrado
Laboratoris Clínics, Hospital Universitari Vall d'Hebron, Barcelona

1. Introducció

Les *variables discretes* són aquelles que només poden prendre certs valors determinats pertanyents a un conjunt numerable, per contraposició amb les *variables contínues,* que poden prendre qualsevol valor dins d'un interval determinat (1). Per exemple, el nombre de fills d'una parella és una variable discreta que pot ser 0, 1, 2,... però mai podrà ser 1,4. En canvi l'alçada d'una persona pot prendre qualsevol valor dins d'un rang (el de les alçades possibles pels éssers humans); 1,62 m, 1,85 m, 2,06 m, etc.

Els valors de les variables discretes poden pertànyer a escales binàries o dicotòmiques, amb dos valors possibles, com per exemple la presència o no per sobre d'una concentració determinada d'una droga d'abús en l'orina; o a escales polinàries o politòmiques, amb més de dos valors possibles, com en l'exemple anterior el nombre de fills (2, 3).

Dins les propietats estudiades al laboratori clínic, les que poden assimilar-se a variables discretes pel seu estudi estadístic són les propietats qualitatives, per exemple el grup sanguini (propietat qualitativa polinària) o el sexe (propietat qualitativa binària); i les propietats ordinals (o magnituds ordinals), com per exemple la concentració de bacteris en un sediment urinari expressat com a absent, escàs, moderat o abundant.

[1] Membres de la Secció d'Estadística i Metrologia durant la preparació d'aquest document: A. Blanco Font, B. Candás Estébanez, X. Fuentes Arderiu (president), M. Martínez Casademont, M. Mosquera Parrado, J.M. Queraltó Compañó, L. Rami Brualla, R. Rigo Bonnin, H. Valbuena Parralejo.

Les variables discretes no permeten cap transformació matemàtica, tret del recompte i transformacions monòtones en el cas de les variables discretes assimilables a propietats ordinals, i els estadístics que s'acostumen a calcular per a aquest tipus de variable són la moda i algun índex de dispersió. També permeten l'aplicació de l'estadística de proporcions. Per tant, el "joc estadístic" que permeten és més limitat que en el cas de les variables contínues.

L'objectiu d'aquest text és realitzar una revisió de les bases de l'estadística descriptiva aplicada a variables discretes.

Estadística descriptiva aplicada a variables discretes

En l'estudi estadístic descriptiu d'una variable discreta, ja sigui binària o polinària, les dades obtingudes es recullen habitualment en una *taula de freqüències*. Aquestes taules es poden representar de manera gràfica amb *diagrames de barres* o *gràfiques de sectors*. Per últim, com ja s'ha indicat, la *moda* i l'índex de dispersió són els únics estadístics que s'acostumen a aplicar quan s'estudien aquests tipus de variables.

discretes.

Taules de freqüència (1, 4, 5)

Es pren una mostra de grandària *n* en la què es vol estudiar una certa variable discreta amb un nombre de valors possibles i determinats k.

Es defineix la *freqüència absoluta* d'un valor i, representada com a n_i, com el nombre d'individus dins de la mostra que presenten aquest valor i. La suma de les freqüències absolutes de tots els valors de la mostra ha de ser igual a n.

$$\sum_{i=1}^{i=k} n_i = n_1 + n_2 + \ldots + n_k = n$$

La *freqüència relativa* d'un valor i, representada com a f_i, es calcula dividint la freqüència absoluta d'un valor *i* en la mostra entre el nombre total d'individus, *n*. La suma de les freqüències relatives de tots els valors de la mostra ha de ser per tant igual a 1; si les freqüències relatives s'expressen com a percentatge, la suma haurà de ser igual a 100 (Taula 1).

$$fi = \frac{n_i}{n}, \qquad \sum_{i=1}^{i=k} f_i = f_1 + f_2 + \ldots + f_k = 1$$

Les freqüències relatives també reben el nom de *proporcions* (símbol: *p*).

Valor	Freqüència absoluta	Freqüència relativa
1	n_1	f_1
2	n_2	f_2
...	...	...
K	n_k	f_k
Total	N	1

Taula 1. Freqüències absolutes i relatives.

<u>Exemple 1. Variable discreta binària corresponent a una magnitud ordinal</u>

S'estudia una mostra de 100 pacients del Servei d'Urgències a qui se'ls demana la mesura de Uri—Coriogonadotropina; c.arb.({negatiu; positiu}), popularment anomenada *prova d'embaràs*, i el valor mesurat és *positiu* en 12 de les pacients. Aquestes proves d'embaràs es basen en la mesura de la concentració de coriogonadotropina en l'orina; qualsevol valor mesurat de la concentració d'aquesta hormona en l'orina superior al límit de detecció del sistema de mesura emprat es considera positiu, la qual cosa indica que existeix compatibilitat biològica entre el valor mesurat i el possible embaràs de la pacient. Cal ressaltar que un valor mesurat negatiu no vol dir que en la mostra d'orina es descarti la presència de coriogonadotropina; un valor mesurat negatiu només ens assegura que la concentració de coriogonadotropina es troba per sota del límit de detecció del sistema de mesura emprat.

En aquest exemple, la grandària de la mostra és 100 pacients, els valors possibles de la variable considerada són positiu i negatiu i les freqüències absolutes dels resultats trobats són 12 positius i 88 negatius, corresponents a les freqüències relatives 0,12 (12 %) positius i 0,88 (88 %) negatius.

A la Taula 2 es mostra la taula de freqüències per aquest exemple.

Uri—Coriogonadotropina; c.arb.({negatiu; positiu})	**Freqüència absoluta**	**Freqüència relativa**
Positiu	12	0,12 (12 %)
Negatiu	88	0,88 (88 %)
Total	100	1 (100 %)

Taula 2. Exemple 1. Variable discreta binària corresponent a una magnitud ordinal.

<u>Exemple 2. Variable discreta binària corresponent a una propietat qualitativa</u>

En el laboratori clínic d'un hospital es realitzen cariotips a partir de mostres de biòpsies corials o de líquids amniòtics per a la detecció prenatal de cromosomopaties. Les cromosomopaties es divideixen en dos grups, estructurals i numèriques. La síndrome de Down és en el 95% dels casos descrits una aneuploïdia numèrica que correspon a la trisomia del cromosoma 21.

Des del punt de vista metrològic (*lato sensu*), la propietat biològica identificada és: Cariotip—Cromosoma 21; trisomia({negatiu, positiu}).

Durant un any, el laboratori clínic ha estudiat 535 cariotips per la detecció prenatal de cromosomopaties i ha detectat 11 casos en els que el cariotip presentava una trisomia del cromosoma 21.

A la Taula 3 es mostra la taula de freqüències per a aquest exemple.

Cariotip—Cromosoma 21; trisomia({negatiu, positiu})	**Freqüència absoluta**	**Freqüència relativa**
Positiu	11	0,021 (2,1 %)
Negatiu	524	0,979 (97,9 %)
Total	534	1 (100 %)

Taula 3. Exemple 2. Variable discreta binària corresponent a una propietat qualitativa.

<u>Exemple 3. Variable discreta polinària corresponent a una propietat qualitativa</u>

El laboratori clínic d'un hospital identifica els grups sanguinis AB0 i Rh_0. Des del punt de vista metrològic (*lato sensu*), la propietat biològica en estudi és: Ers(San)—Antígens eritrocítics; taxonalitat(ABO; Rh). En 157 mostres s'obtenen els resultats detallats a la Taula 4.

Ers(San)—Antígens eritrocítics; taxonalitat(ABO; Rh D)	**Freqüència absoluta**	**Freqüència relativa**
O, Rh positiu	57	0,363 (36,3 %)
A, Rh positiu	53	0,338 (33,8 %)
B, Rh positiu	12	0,076 (7,6 %)
AB, Rh positiu	4	0,025 (2,5 %)
O, Rh negatiu	14	0,089 (8,9 %)
A, Rh negatiu	12	0,078 (7,6 %)
B, Rh negatiu	3	0,019 (1,9 %)
AB, Rh negatiu	2	0,013 (1,3 %)
Total	157	1 (100,0 %)

Taula 4. Exemple 3. Variable discreta polinària corresponent a una propietat qualitativa.

Només en el cas que s'estudiï una variable discreta corresponent a una magnitud ordinal amb una escala polinària de valors possibles, és interessant calcular també la *freqüència acumulada* de cadascun dels valors. La freqüència acumulada reflectirà directament el nombre d'individus de la mostra que presenten un valor inferior de la variable d'estudi a un cert valor; per aquest motiu no té sentit el seu càlcul per a variables discretes corresponents a propietats qualitatives.

La *freqüència absoluta acumulada* per a un valor i, N_i, s'obté sumant a la freqüència absoluta del valor i les freqüències absolutes de tots els valors de la taula de freqüències inferiors a aquest valor i. La freqüència absoluta acumulada de l'últim valor (N_k) de la taula és, per tant, igual a n.

$$N_i = \sum_{j=1}^{j=i} n_j = n_1 + n_2 + \ldots + n_i$$

$$N_k = \sum_{j=1}^{j=k} n_j = n_1 + n_2 + \ldots + n_k = n$$

On i és el valor pel qual es vol calcular la freqüència absoluta acumulada i k és el nombre total de valors.

Igualment, la *freqüència relativa acumulada per a un valor* i, F_i, s'obté sumant a la freqüència relativa del valor i, les freqüències relatives dels valors inferiors. La freqüència relativa acumulada de l'últim valor de la taula és, per tant, igual a 1.

$$F_i = \sum_{j=1}^{j=i} f_j = f_1 + f_2 + \ldots + f_i$$

$$F_k = \sum_{j=1}^{j=k} f_j = f_1 + f_2 + \ldots + f_k = 1$$

<u>Exemple 4. Variable discreta polinària corresponent a una magnitud ordinal</u>

Es pren una mostra de 40 sediments que pertanyen a 40 pacients que arriben al Servei d'Urgències d'un hospital amb simptomatologia d'infecció del tracte urinari. El conjunt de valors possibles de concentració arbitrària de bacteris en el sediment urinari establert al laboratori clínic de l'hospital que es tracta és {absents, escassos, bastants, abundants}.

Des del punt de vista metrològic, la magnitud biològica ordinal en estudi és: Uri—Bacteris; c.arb.(sediment; microscòpia; {absents, escassos, bastants, abundants})

A la Taula 5 es mostren els resultats d'aquest estudi, ordenant els valors segons la quantia que denoten. S'han inclòs també les freqüències acumulades, tant absolutes com relatives.

Concentració arbitrària de bacteris en l'orina	Freqüència absoluta	Freqüència relativa (%)	Freqüència absoluta acumulada (N_i)	Freqüència relativa acumulada (F_i)
Absents	1	0,025 (2,5 %)	1	0,025 (2,5 %)
Escassos	5	0,125 (12,5 %)	6 (5+1)	0,150 (15 %)
Bastants	16	0,400 (40 %)	22 (16+5+1)	0,550 (55 %)
Abundants	18	0,450 (45 %)	40 (18+16+5+1)	1 (100 %)
Total	40	1 (100 %)		

Taula 5. Exemple 4. Variable discreta polinària corresponent a una magnitud ordinal.

Interpretant aquesta taula es pot dir, per exemple, que 22 dels 40 pacients (el 55%) tenien una concentració arbitrària de bacteris en l'orina menor o igual al valor «bastants», o que només 6 pacients (15%) tenien una concentració arbitrària de bacteris en l'orina igual o inferior al valor «escassos».

Representacions gràfiques de les taules de freqüències (1, 5)

En l'estudi estadístic d'una variable discreta, les representacions gràfiques de les dades es basen en la proporcionalitat de les freqüències absolutes o relatives dels valors a una àrea.

Es poden trobar principalment dos tipus de representació: els diagrames de barres i els diagrames de sectors.

Un *diagrama de barres* està constituït per tants rectangles com valors presenti la variable en la mostra estudiada. Tots els rectangles tindran una base d'igual amplitud i l'altura serà proporcional a la freqüència absoluta o relativa; així l'àrea del rectangle és proporcional a la freqüència que representa.

A les figures 1 i 2 es mostren les representacions dels diagrames de barres corresponents als exemples 1 i 4.

Figura 1. Diagrama de barres de l'exemple 1.

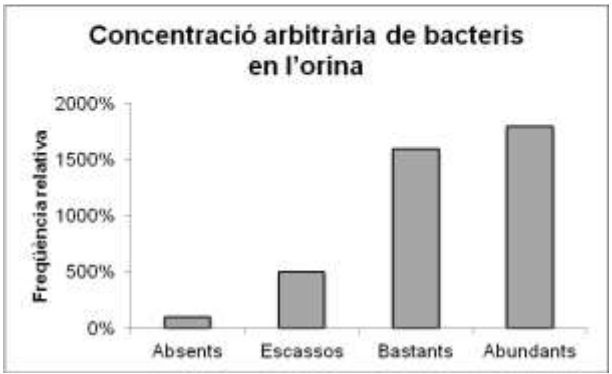

Figura 2. Diagrama de barres de l'exemple 4.

Un *diagrama de sectors* es construeix dividint un cercle en tants sectors com valors presenti la variable en estudi. L'angle de cada sector, i conseqüentment l'àrea de cada sector, és proporcional a la freqüència del valor en la mostra en estudi, tant absoluta com relativa. Es poden acompanyar d'una llegenda de colors que explica a quin valor correspon cada sector.

A les figures 3 i 4 es mostren les representacions dels diagrames de sectors corresponents als exemples 2 i 3.

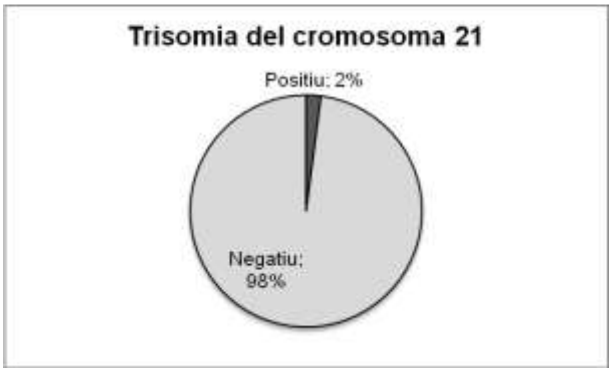

Figura 3. Diagrama de sectors de l'exemple 2.

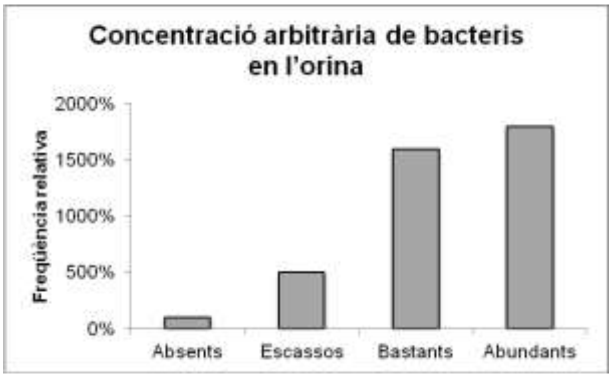

Figura 4. Diagrama de sectors de l'exemple 3.

Estadístics de tendència central (1)

La *moda* d'una mostra o població és el valor amb una freqüència (absoluta o relativa) o proporció més gran. Com ja s'ha dit, és l'únic estadístic de tendència central que es pot aplicar a les variables discretes.

Es pot donar el cas de tenir una mostra bimodal, trimodal, etc. si dos o més valors diferents tenen freqüències iguals i són les més grans de la mostra.

La moda en l'exemple 1, on s'estudiava la concentració arbitrària de l'hormona coriogonadotropina en l'orina és el valor «negatiu».

En el cas de l'exemple 2, detecció prenatal de la trisomia del cromosoma 21 en el cariotip, la moda també és el valor «negatiu». En l'exemple 3, el grup sanguini amb una freqüència més elevada és el «grup O, Rh positiu» i per tant n'és la moda. Per últim, en l'exemple 4, on s'estudiava la concentració arbitrària de bacteris en l'orina, la moda és el valor «abundants».

Estadístics de dispersió

A diferència de la importància que se li atorga a la dispersió quan s'estudien variables contínues, la dispersió aplicada a les variables discretes no és freqüentment tractada als llibres d'estadística. Aquest fet no és degut a la falta d'eines per estimar aquesta dispersió, sinó més aviat a una manca de consens dels experts per decidir quina aproximació, de les moltes proposades, és la més correcta i útil.

Quan es treballa amb variables quantitatives, la forma més habitual d'estudiar la dispersió de les dades consisteix en calcular la variació que existeix entre les dades d'una mostra i la seva mitjana. Però en el cas de les variables discretes, aquest plantejament no és possible, ja que no se'n pot calcular la mitjana.

Per a variables discretes s'ha d'entendre la variabilitat d'una manera alternativa; en comptes d'estudiar la variació entre totes les dades d'una mostra i la seva mitjana, es poden estudiar la variació de les dades de la mostra entre sí.

En aquesta part d'aquest document docent es presentaran algunes d'aquestes propostes per al càlcul de diversos estadístics de dispersió per a variables discretes i es discutirà breument la seva aplicació.

Raó de variació

La raó de variació és un estadístic de dispersió per a variables discretes de tota mena molt senzill de calcular (6). El seu valor és la proporció d'individus d'una mostra que, per a la propietat biològica en estudi, tenen un valor diferent al de la moda.

En una mostra de grandària n amb un nombre d'individus n_{mo} que, per la propietat en estudi, presenten el valor de la moda, la raó de variació (v:) es calcula de la següent manera:

$$v := 1 - \frac{n_{mo}}{n}$$

El valor de la raó de variació serà igual a 0 si tots els individus de la mostra presenten el mateix valor i s'allunyarà més d'aquest valor com més repartits estiguin els individus entre els valors possibles, és a dir, quanta més variabilitat presenti la mostra.

La raó de variació no considera el nombre de valors possibles ni el nombre d'individus repartits entre aquests valors, per tant la informació que se n'extreu és més limitada que amb altres estadístics de dispersió presentats en aquest apartat.

<u>Càlcul de la raó de variació per a les dades de l'exemple 2.</u>

La moda en aquest exemple és el valor «negatiu».

$$v := 1 - \frac{n_{mo}}{n} = 1 - \frac{88}{100} = 0{,}12$$

<u>Càlcul de la raó de variació per a les dades de l'exemple 4.</u>

La moda en aquest exemple és el valor «abundants».

$$v := 1 - \frac{n_{mo}}{n} = 1 - \frac{18}{40} = 0{,}55$$

Índex de variació qualitativa

L'índex de variació qualitativa (d'ara endavant *IVQ*) és un estimador de la variabilitat basat en el càlcul del quocient entre el nombre de diferències observades en les dades i el nombre de diferències màxim possible. S'han proposat equacions diferents per calcular-lo (7, 8, 9) que aplicades a una mateixa mostra proporcionarien valors diferents. Tot i així, totes aquestes alternatives compleixen els requisits següents:

- El resultat de l'*IVQ* varia entre 0 i 1.
- L'*IVQ* és 0 en el cas que tots els individus de la mostra tinguin el mateix valor i aquest sigui l'únic valor possible.
- L'*IVQ* és 1 en el cas que els individus de la mostra estiguin igualment distribuïts entre tots els valors *k* possibles.
- Si es calcula l'*IVQ* amb la mateixa equació en dues mostres diferents on s'estudia la mateixa variable discreta, la mostra amb l'*IVQ* més proper a 1 és la que presenta una major dispersió de les dades.

Una de les propostes més emprades pel càlcul de l'*IVQ* és la de Gibbs *et al.* 1975 (7). En una mostra on es vol estudiar una variable discreta amb *k* valors possibles, l'*IVQ* es calcula de la manera següent:

$$IVQ = \frac{k}{k-1}\left(1-\sum_{i=1}^{k} f_i^2\right)$$

on f_i són les freqüències relatives dels valors a la mostra.

L'*IVQ* es considera actualment una bona aproximació al càlcul de la variació per a les mostres on s'estudia una variable discreta que correspon a una propietat qualitativa, però <u>no és apropiada per a magnituds ordinals</u> (10). L'*IVQ* tracta totes les diferències entre els valors possibles de la variable per igual. Això és correcte per a una variable discreta que correspon a una propietat qualitativa; per exemple per al color d'un líquid peritoneal, on vermellós és tan diferent de groc com ho és de verdós. Però, en canvi, els valors d'una magnitud ordinal denoten una quantia i per tant les diferències entre els diferents valors no són les mateixes; en l'estudi d'un sediment d'orina no és igual la diferència entre una concentració arbitrària de bacteris «abundant» o «bastant» que entre una concentració «abundant» o «absent».

<u>Càlcul de l'*IVQ* amb l'equació proposada per Gibbs *et al.* per a l'exemple 2.</u>

Les dades es troben en la taula 3.

$$IVQ = \frac{k}{k-1}\left(1-\sum_{i=1}^{k} f_i^2\right) = \frac{2}{2-1}\left[1-\left(0{,}021^2+0{,}979^2\right)\right] = 0{,}082$$

En aquest exemple l'*IVQ* és molt proper a 0, ja que el 97,9 % dels resultats de la mostra corresponents a la propietat biològica estudiada, Cariotip—Cromosoma 21; trisomia({negatiu, positiu}), eren negatius.

<u>Càlcul de l'*IVQ* amb l'equació proposada per Gibbs *et al.* per a l'exemple 3.</u>

La propietat biològica estudiada en aquest exemple és Ers(San)—Antígens eritrocítics; taxonalitat (ABO; Rh), les dades es troben a la Taula 4.

$$IVQ = \frac{k}{k-1}\left(1-\sum_{i=1}^{k} f_i^2\right) =$$
$$\frac{8}{8-1}\left[1-\left(0{,}363^2+0{,}338^2+0{,}076^2+0{,}025^2+0{,}089^2++0{,}019^2+0{,}013^2\right)\right] = 0{,}84$$

Encara que els *IVQ* calculats per a variables diferents i amb escales diferents no són comparables, podem veure que el valor de l'*IVQ* en aquest exemple és més gran que en l'anterior i que efectivament això es relaciona amb una distribució més uniforme dels resultats entre els valors possibles.

Entropia de dispersió

L'entropia de dispersió s'utilitza com a estimador de la variabilitat de les variables discretes corresponents a propietats qualitatives (nominals) (11). L'entropia de dispersió (H) d'una mostra amb k categories possibles és igual a la suma negativa dels logaritmes neperians de les freqüències relatives per a cada valor de la variable, multiplicats per la mateixa freqüència relativa:

$$H = -\sum_{i=1}^{k} f_i \times \ln(f_i)$$

El valor de l'entropia de dispersió presenta un valor mínim de 0, quan tots els individus d'una mostra presenten el mateix valor de la variable.

$$H = -\sum_{i=1}^{k} f_i \times \ln(f_i) = -1 \times \ln(1) = 0$$

El valor màxim de l'entropia de dispersió depèn del nombre de valors possibles de la variable discreta, i es dóna quan tots els individus es reparteixen per igual entre tots els valors possibles. Per exemple, en una mostra amb k=4, on tots els individus es reparteixen de manera equitativa entre els k valors possibles, l'entropia de dispersió màxima serà:

$$H = -\sum_{i=1}^{k} f_i \times \ln(f_i) = -4[0{,}25 \times \ln(0{,}25)] = 1{,}39$$

Quan es compara l'entropia de dispersió de dues mostres diferents amb el mateix nombre de valors de la variable discreta en estudi, aquella que presenti un valor de l'entropia de dispersió més gran, és la que té major variabilitat.

L'entropia de dispersió suposa una estratègia alternativa a *l'IVQ* pel càlcul de la dispersió en mostres on s'estudia una propietat qualitativa. Dóna una informació complementaria, però no presenta equivalències directes amb les fórmules de l'*IVQ*.

Pel mateix motiu que l'*IVQ,* l'entropia de dispersió <u>no és apropiada per a magnituds ordinals,</u> ja que omet en el seu càlcul que els valors d'una magnitud ordinal denoten una quantia i que per tant no tots els valors són igual de diferents entre si.

El quocient entre l'entropia de dispersió (H) d'una mostra i la màxima possible ($H_{màx}$) és

l'anomenat *índex d'uniformitat de Pielou* (J) (12):

$$J = \frac{H}{H_{màx}}$$

El valor màxim de l'índex d'uniformitat de Pielou és 1, i es donarà quan l'entropia de dispersió d'una mostra sigui la màxima, és a dir, quan la variabilitat és la màxima perquè tots els individus estan uniformement repartits entre tots els valors possibles. En canvi l'índex d'uniformitat de Pielou serà 0 quan l'entropia de dispersió sigui la mínima; és en el cas que tots els individus de la mostra presentin el mateix valor.

<u>Càlcul de l'entropia de dispersió per a l'exemple 2.</u>

Les dades es troben en la Taula 3.

$$H = -\sum_{i=1}^{k} f_i \times \ln(f_i) = -[0{,}021 \times \ln(0{,}021)) - (0{,}979 \times \ln(0{,}979)] = 0{,}1$$

L'entropia de dispersió màxima i l'índex d'uniformitat de Pielou en aquest cas seran:

$$H_{\max} = -\sum_{i=1}^{k} f_i \times \ln(f_i) = -2 \times [0{,}5 \times \ln(0{,}5)] = 0{,}69$$

$$J = \frac{0{,}1}{0{,}69} = 0{,}14$$

<u>Càlcul de l'entropia de dispersió per l'exemple 3.</u>

Les dades es troben en la Taula 4.

$$H = -\sum_{i=1}^{k} f_i \times \ln(f_i) = -[0{,}338 \times \ln(0{,}338)] - \ldots = 1{,}57$$

L'entropia de dispersió màxima i l'índex d'uniformitat de Pielou en aquest cas seran:

$$H_{max} = -\sum_{i=1}^{k} f_i \times \ln(f_i) = -8 \times [0{,}152 \times \ln(0{,}125)] = 2{,}08$$

$$J = \frac{1{,}57}{2{,}08} = 0{,}75$$

Índex de Simpson/ Herfindahl-Hirschman

L'índex de Simpson/Herfindahl-Hirschman (λ) per a variables discretes corresponents a propietats qualitatives (nominals) es calcula com la suma quadràtica de les proporcions (p), o freqüències relatives, observades per a cada valor de la variable en estudi (13):

$$\lambda = \sum_{i=1}^{n} p_i^2$$

El valor màxim d'aquest índex és 1, que correspon a la mínima variabilitat, i es produeix quan tots els individus d'una mostra presenten el mateix valor. El valor mínim és $1/k$, on k és el nombre de valors possibles de la variable, i es produeix quan la dispersió és màxima, és a dir, els individus es distribueixen de manera equitativa entre tots els valors.

Igual que l'entropia de dispersió o que l'*IVQ*, aquest índex <u>no és apropiat per a l'estudi de magnituds ordinals.</u>

<u>Càlcul de l'índex de Simpson/Herfindahl-Hirschman per a l'exemple 3.</u>

Les dades es troben en la Taula 4.

$$\lambda = \sum_{i=1}^{n} p_i^2 = (0{,}363^2 + 0{,}338^2 + \ldots) = 0{,}267$$

En aquest cas, el valor mínim (màxima variabilitat) de l'índex és:

$$\lambda = \sum_{i=1}^{n} p_i^2 = 8 \times 0{,}125^2 = \frac{1}{8} = 0{,}125$$

Estimadors de la variabilitat per a les magnituds ordinals

Ja s'ha dit que l'aplicació de l'*IVQ* o de l'entropia de dispersió no resulta apropiada quan es treballa amb variables discretes corresponents a una magnitud ordinal. Existeixen diverses propostes publicades per respondre a la necessitat d'un estimador de dispersió per aquest cas (10, 14, 15). D'entre elles, una de les més acceptades és la proposada per Blair i Lavy en la qual es calcula un estimador de dispersió anomenat h^2:

$$h^2 = \frac{1}{(k-1)/4} \times \sum_{i=1}^{i=k-1} F_i \times (1 - F_i)$$

On k és el nombre de valors possibles de la variable i F_i és la freqüència acumulada per cadascun dels valors.

Els valors de h^2 oscil·len entre 0 i 1. El valor h^2=1 correspon a la situació en que els individus de la mostra estudiada es reparteixen per igual entre els dos valors extrems de la variable (en el cas d'una variable dicotòmica, es repartirien per igual entre els dos valors). Per contra, el valor mínim $h^2=0$ correspon a una mostra amb un únic valor possible per la variable, és a dir k=1. Com en els casos anteriors, no es recomana utilitzar h^2 per comparar la dispersió de dues mostres en les quals s'ha calculat la dispersió en base a escales diferents.

<u>Càlcul d'h^2 per a l'exemple 1.</u>

Les dades es troben en la Taula 2. Com es tracta d'una variable dicotòmica, l'única freqüència acumulada a tenir en compte és la del valor més baix («negatiu»).

$$h^2 = \frac{1}{(k-1)/4} \times \sum_{i=1}^{i=k-1} F_i \times (1 - F_i) = \frac{1}{(2-1)/4} \times 0{,}88 \times (1 - 0{,}88) = 0{,}42$$

Càlcul d'h^2 per a l'exemple 4.

Les dades es troben en la Taula 5.

$$h^2 = \frac{1}{(k-1)/4} \times \sum_{i=1}^{i=k-1} F_i \times (1-F_i) =$$

$$\frac{1}{(4-1)/4} \times [0{,}025 \times (1-0{,}025) + 0{,}15 \times (1-0{,}15) + 0{,}55 \times (1-0{,}55)] = 0{,}6$$

Per acabar aquest apartat val la pena recalcar que, com ja s'ha dit al començament, existeixen altres aproximacions a la qüestió del càlcul d'estadístics de dispersió per a variables discretes, descrites i discutides a la literatura especialitzada. Es recomana consultar la bibliografia en que s'ha basat aquest document docent per aprofundir en aquest tema.

Bibliografia

1. García Pérez A. Estadística aplicada: Conceptos básicos. Madrid: UNED; 2012.

2. Trullols E, Ruisanchez I, Rius FX. Validation of qualitative analytical methods.Trends in Anal Chem 2004;23:137-145.

3. Fuentes Arderiu X, Miró Balagué J. Naturalesa de les propietats biològiques examinades al laboratori clínic *In vitro veritas* 2011;12:150-159 <http://www.acclc.cat/continguts/ivv135.pdf> (Accés 2013-07-05).

4. Doménech Massons J M. Bioestadística: Métodos estadísticos para investigadores. Barcelona: Herder; 1977.

5. Forcada Plaza S, Rubió Massegú J. Elements d'estadística. Barcelona: UPC; 2007.

6. Freeman, Linton C. Elementary applied statistics, New York: John Wiley and Sons; 1965.

7. Light RJ, Margolin BH. An analysis of variance for categorical data. J Am Stat Assoc 1987;66:534-44.

8. Gibbs JP, Poston DL Jr. The Division of labor: Conceptualization and related measures. Social Forces 1975;53:468-76.

9. Wilcox AR. Indices of qualitative variation. Tennessee: Union Carbide Corporation; 1967.

10. Blair J, Lacy MG. Statisticals of ordinal variation. Sociol Method Res 2000;28;251-80.

11. Vanhoy M. An entropy estimator of population variability in nominal data. J Scien Psychol 2008;(June):25-30.

12. Pielou EC. The measurement of diversity in different types of biological collections. J Theoret Biol 1966;13:131-44.

13. Simpson EH. Measurement of diversity. Nature 1949;163:688.

14. Gadrich T, Bahkansky E. Ordanova: Analysis of ordinal variation. J Stat Plan Infer 2012;142:3174-88.

15. Gadrich T, Bahkansky E. Interlaboratory comparison of test results of an ordinal or nominal binary property: analysis of variation. Accr Qual Assur 2012;17:239-43..